# Communicative Facial Displays as a New Conversational Modality

*Akikazu Takeuchi and Katashi Nagao*
Sony Computer Science Laboratory, Inc.
3-14-13 Higashi-Gotanda
Shinagawa-ku, Tokyo 141, Japan
TEL: +81-3-3448-4380
{takeuchi,nagao}@csl.sony.co.jp

## ABSTRACT

The human face is an independent communication channel that conveys emotional and conversational signals encoded as facial displays. Facial displays can be viewed as communicative signals that help coordinate conversation. We are attempting to introduce facial displays into computer-human interaction as a new modality. This will make the interaction tighter and more efficient while lessening the cognitive load. As the first step, a speech dialogue system was selected to investigate the power of communicative facial displays. We analyzed the conversations between users and the speech dialogue system, to which facial displays had been added. We found that conversation with the system featuring facial displays was more successful than that with a system without facial displays.

**KEYWORDS:** User interface design, multimodal interfaces, facial expression, conversational interfaces, anthropomorphism.

## INTRODUCTION AND MOTIVATIONS

In designing computer-human interaction, human face-to-face conversation has provided an ideal model. One of the major features of face-to-face communication is the multiplicity of communication channels. A channel is a communication medium associated with a particular encoding method. Examples are the auditory channel that carries speech, and the visual channel that carries facial expressions. A modality is the sense used to perceive signals from the outside world. One channel may be perceived by more than one sense. The senses of sight, hearing, and touch are all examples of modalities.

Multimodal user interfaces are interfaces with multiple channels that act on multiple modalities. To realize a true multimedia/modal user interface, it is necessary to study how humans perceive information and to which information humans are sensitive.

In usual face-to-face communication, many channels are used and different modalities are activated. Conversation is supported by multiple coordinated activities of various cognitive levels. For instance, syntactic and semantic processing are coupled, and object-level processing (relevant to the communication goal) and meta-level processing (relevant to communication regulations) are executed in parallel as part of these coordinated activities. As a result, communication becomes highly flexible and robust, so that failure of one channel is recovered by another channel, and a message in one channel can be explained by the other channel.

In fact, face-to-face communication is the primary communication style, evolved over aeons, from primates to human beings. Our brain has adapted to this style of communication. As the terms "face-to-face" and "interface" indicate, faces play an essential role in communication. In the field of neurophysiology, it is well-known that a particular region of primates' brains is dedicated to facial information processing [14]. This implies that the ability to process facial information is a major factor in surviving natural selection.

The study of facial expressions has attracted the interest of a number of different disciplines, including psychology, ethology, and interpersonal communication. Facial expressions are viewed in either of two ways. One regards facial expressions as expressions of emotional states [8]. The other views facial expressions in a social context, and regards them as communicative signals [9]. The term "facial displays" is equivalent to "facial expressions", but does not have the connotation of emotion. In this paper, we use the term "facial displays."

The present paper assumes the second view. A face is an independent communication channel that conveys emotional and conversational signals encoded as facial displays. A facial display can be also seen as a modality because the human brain has a special circuit dedicated to the necessary processing. Taking this as a starting point, we are attempting to bring facial displays into computer human interaction as a new modality that makes the interaction tighter and more efficient, while lessening the cognitive load.

Another reason for placing attention on faces is that faces provide a human with social interfaces. As J. S. Brown indicated in the closing talk at CHI'92, future CHI technologies should help people establish and strengthen

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

© 1993 ACM 0-89791-575-5/93/0004/0187...$1.50

their social relationships. Humans are social animals. Facial displays are usually directed not at oneself, but at others. They have been evolved to help us develop better social relations with others. Facial displays are primarily communicative and thus are subject to social factors that regulate their occurrence [2]. Therefore, the study of facial displays is expected to reveal important characteristics of social interfaces.

## RELATED WORK

Our approach is classified into the so-called "multimodal interfaces." Blattner surveyed a range of models to design a multimodal interfaces [1]. Among them, the conversational model and the anthropomorphic models are closely related to our approach.

The conversational model places greater emphasis on conversation structure [10]. In communication, different communication channels are used and several modalities will be activated. A conversation structure exists throughout all of these and helps to coordinate the conversation. To clarify and abstract the (possibly linguistic) structure of conversation is the goal of the conversational model. Our approach focuses on a specific communication channel and modality, namely, human faces. Through evolution, humans developed sophisticated communication using faces. Knowing about the roles of communicative displays leads us to an understanding of how people coordinate conversation by sending signals through multimodal channels, that is, understanding the dynamics of conversation. By applying that knowledge, we will be able to realize a new user interface that can exchange even subtle information using sophisticated communicative displays.

The anthropomorphic model has been a controversial topic in recent CHI conferences [4]. While there are some ad hoc anthropomorphic interfaces, carefully designed interfaces such as GUIDES illustrate the possibilities of new user interfaces [5]. This research is an attempt to computationally capture the communicative power of human faces and to apply it to user interfaces. The resulting interface will, of course, be anthropomorphic. However, we use faces as an additional communication channel connected to an unexplored modality, rather than as a humanity option for the interfaces.

## STUDIES OF FACIAL DISPLAYS

Facial displays have been the subject of scientific study for a long time. Darwin identified two aspects of facial expression; that related to emotion, and that related to communication [3]. We will give a brief introduction to the theory of communicative facial displays below.

### Theory of Communicative Facial Displays

Fridlund and Gilbert [9] proposed that the primary role of facial displays is to provide information that augments the verbal component of communication, rather than to provide emotional information. The theory makes two major assumptions.

The first is that facial displays are primarily communicative. They are used to convey information to other people. The information that is conveyed may be emotional information, or other kinds of information, for example, syntactical information, indications that the speaker is being understood, relationship definition, listener responses, etc. [2]. Facial displays can function in an interaction as communication on their own. That is, they can send a message independently of other communicative behavior. Facial emblems such as winks, facial shrugs, and listener's comments (agreement or disagreement, disbelief or surprise) are typical examples. Facial displays can also work in conjunction with other communicative behavior (both verbal and nonverbal) to provide information. For example, facial displays in social interaction may function to reduce ambiguity in spoken language, as stress patterns, voice tone, and kinetic behavior also act in the same way. This is done by illustrating or adding information, or by meta communication, that is, communication about how a message should be taken. (e.g., smiling when joking).

The second assumption is that facial displays are primarily social. They occur for the purpose of communicating information to others. Their occurrence is regulated more by the social situation than by any underlying emotion processes.

### Facial Displays in Conversation

There have been several attempts to categorize facial displays according to their communicative roles [2,6,15]. In a similar vein, the categorization of emotional facial displays by Ekman and Friesen [8] is well known. Emotional facial displays are basically independent of the situation, that is, their meanings are the same wherever and whenever they appear. Unlike the emotional categorization, in communicative categorization almost all displays, except for those displays known as facial emblems, are situation dependent. Namely, the interpretation of a communicative facial display depends upon the situation in which it appears. Conversational context and chronological relations (e.g., synchronization or delay) with other communication channels have a significant impact in their interpretations.

Table 1 is assembled from previous work on categorizing communicative facial displays [2,6,15]. The table lists three major categories:

Syntactic Displays. These are defined as facial displays that (1) mark stress on particular words or clauses, (2) are connected with the syntactic aspects of an utterance or (3) are connected with the organization of the talk.

Speaker Displays. Speaker displays are defined as being facial displays that (1) illustrate the idea being verbally conveyed, or (2) add additional information to the ongoing verbal content.

Listener Comment Displays. These are facial displays made by the person who is not currently speaking and which are made in response to the utterances of the other person.

## SPEECH DIALOGUE WITH FACIAL DISPLAYS

As a first step, a speech dialogue system was selected to investigate the power of communicative facial displays.

### Prototype Architecture

The system consists of two subsystems, a facial animation subsystem that generates a three-dimensional face capable of facial display, and a speech dialogue subsystem that recognizes and interprets speech, and generates voice output. Currently, the animation subsystem is running on an SGI 320VGX, and the speech dialogue subsystem on a Sony NEWS workstation. These two subsystems communicate with each other via an Ethernet network. Figure 1 illustrates the system architecture.

### Facial Animation Subsystem

The face is modeled three-dimensionally. The face is composed of approximately 500 polygons. The face may be rendered using a skin-like surface material with Gouraud shading or by applying a texture map taken from a video frame or a picture.

Table 1: Communicative Categorization of Facial Displays

| | |
|---|---|
| SYNTACTIC DISPLAYS | |
| 1. Exclamation marks | Eyebrow raising |
| 2. Question marks | Eyebrow raising or lowering |
| 3. Emphasizers | Eyebrow raising or lowering |
| 4. Underliners | Longer eyebrow raising |
| 5. Punctuations | Eyebrow movements |
| 6. End of an utterance | Eyebrow raising |
| 7. Beginning of a story | Eyebrow raising |
| 8. Story continuation | Avoid eye contact |
| 9. End of a story | Eye contact |
| SPEAKER DISPLAYS | |
| 10. Thinking/Remembering | Eyebrow raising or lowering,<br>Closing the eyes,<br>Pulling back one mouth side |
| 11. Facial shrug/ "I don't know" | Eyebrow flashes,<br>Mouth corners pulled down,<br>Mouth corners pulled back |
| 12. Interactive/ "You know?" | Eyebrow raising |
| 13. Metacommunicative/ Indication of sarcasm or joke | Eyebrow raising and looking up and off |
| 14. "Yes" | Eyebrow actions |
| 15. "No" | Eyebrow actions |
| 15. "Not" | Eyebrow actions |
| 17. "But" | Eyebrow actions |
| LISTENER COMMENT DISPLAYS | |
| 18. Backchannel/ Indication of attendance | Eyebrow raising,<br>Mouth corners turned down |
| 19. Indication of loudness | Eyebrows drawn to center |
| Understanding levels | |
| 20. Confident | Eyebrow raising, Head nod |
| 21. Moderately confident | Eyebrow raising |
| 22. Not confident | Eyebrow lowering |
| 23. "Yes" | Eyebrow raising |
| Evaluation of utterance | |
| 24. Agreement | Eyebrow raising |
| 25. Request for more Info | Eyebrow raising |
| 26. Incredulity | Longer eyebrow raising |

In 3D computer graphics, a facial display is realized by local deformation of the polygons representing the face. Waters showed that deformation that simulates an action of the muscles underlying the face looks more natural [18]. Therefore, we used the numerical equations simulating muscle actions as defined by Waters. Currently, the system incorporates 16 muscles and 10 parameters controlling mouth opening, jaw rotation, eye movement, eyelid opening, and head orientation. Waters determined these 16 muscles by considering the correspondence with the action units in Facial Action Coding System (FACS) [7]. For details of the facial modeling and animation system, see [16].

Takeuchi, Nagao, Color Plate 1 shows 26 synthesized facial displays corresponding to those listed in Table 1, as well as two additional displays. All facial displays are generated by the method described above, and rendered with a texture map of a young boy. The additional displays are "smile" and "neutral." The neutral display features no muscle contraction, and is used when no conversational signal is needed.

At run-time, the animation subsystem awaits a request from the speech subsystem. When the animation subsystem receives a request that specifies values for the 26 parameters, it starts to deform the face using the received values. The deformation process is controlled by the following differential equation:

$$\dot{f} = a - f$$

where $f$ is a parameter value at time $t$ and $\dot{f}$ is its time derivative at time $t$. $a$ is the target value specified in the request. Using this equation, deformation is fast in the early phase, slowing in the later phase to mimic the real dynamics of facial displays. Currently, the base performance of the animation subsystem ranges from 20-25 frames per second on an SGI Power Series. This is satisfactory for real-time animation.

### Speech Dialogue Subsystem

Our speech dialogue subsystem works as follows: First, a voice input is acoustically analyzed by a built-in sound processing board. Then, a speech recognition module is invoked and outputs word sequences that are assigned higher scores by a probabilistic phone model. These word sequences are syntactically and semantically analyzed and disambiguated by using a relatively loose grammar and a restricted domain knowledge. From the semantic representation of the input utterance, a plan recognition module extracts the speaker's intention. For example, from the utterance "I am interested in Sony's workstation," the module interprets the speaker's intention as "he wants to get precise information about Sony workstation." Once the system determines the speaker's intention, the response generation module is invoked. It generates a system response that satisfies the speaker's intention. Finally, the system's response is output as voice by a voice synthesis module.

Each module, except the voice synthesis module, can send messages to the facial animation subsystem about which facial display should be generated. The correspondence between the situation of speech dialogue and facial displays are listed in Table 2.

The task of the system is to provide information about Sony's computer-related products. For example, the system can answer questions about price, size, weight, and the specifications of Sony's workstations and PCs.

Next, we will present a more detailed description of modules in the speech dialogue subsystem.

*Speech recognition.* This module was developed in cooperation with Tokyo Institute of Technology. Speaker-independent continuous speech input is accepted without special hardware. To obtain a high degree of accuracy, the context-dependent phone Hidden Markov Models are used to construct phone-level hypotheses [11]. This module can generate N-best word-level hypotheses.

*Syntactic and semantic analysis.* This module consists of a parsing mechanism, a semantic analyzer, a relatively loose grammar that contains 24 rules, a lexicon that includes 34 nouns, 8 verbs, 4 adjectives and 22 particles, and a frame-based knowledge base with 61 conceptual frames. Our parsing mechanism is based on Tomita's generalized LR parsing method [17]. Our semantic analyzer can disambiguate ambiguous syntactic structures and generate a semantic representation of the speaker's utterance [12].

*Plan recognition.* This module determines the speaker's intention by constructing his belief model and dynamically adjusting and expanding the model as the dialogue progresses [13]. The module also maintains the topic of the current dialogue and resolve anaphora (reference of pronouns) and ellipsis (omission of subjects).

*Response generation.* This module generates a response by using domain knowledge (database) and text templates (typical patterns of utterances). It selects appropriate templates and combines them to construct a response that satisfies the speaker's intention.

*Correspondence between conversational situations and facial displays.* The speech dialogue subsystem recognizes a number of typical conversational situations that are important in dialogue. We associate these situations with facial displays. For example, in situations where speech inputs are not recognized or where they are syntactically

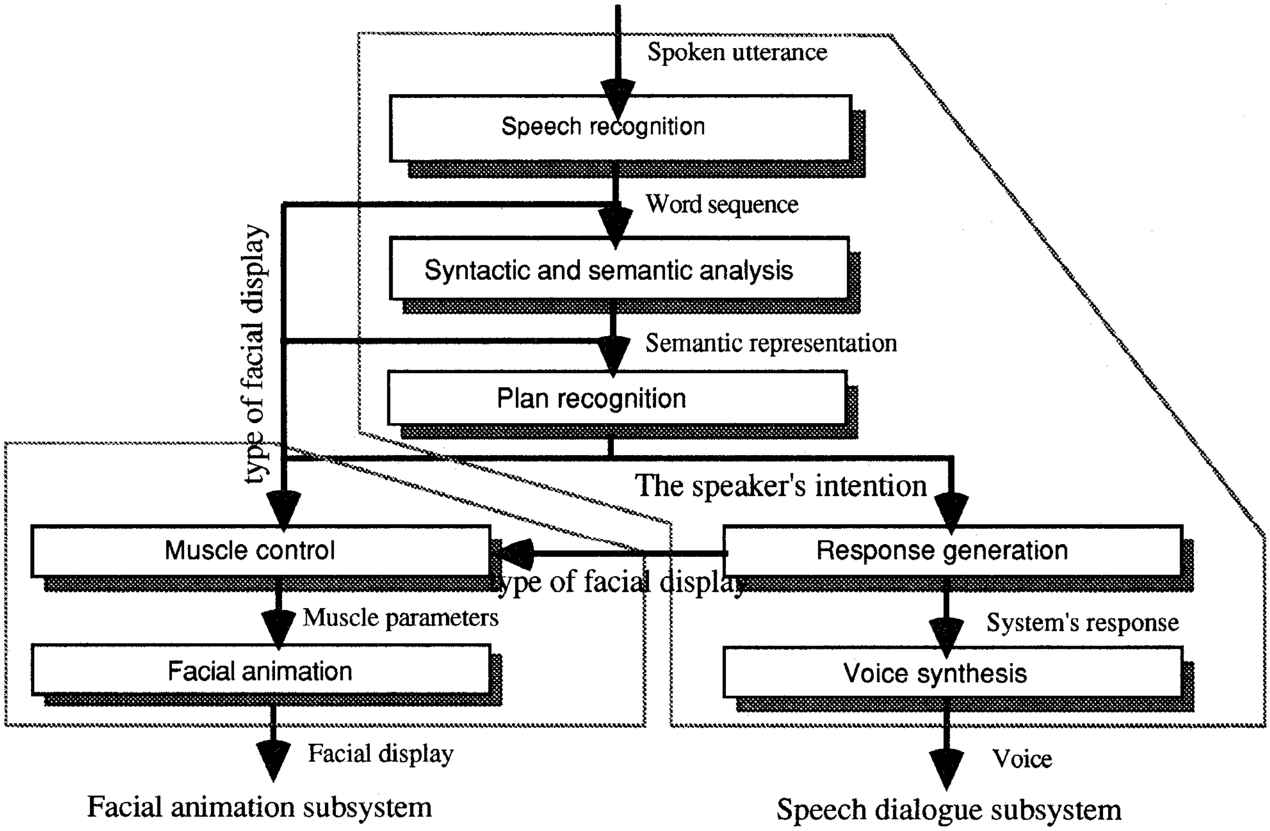


Figure 1. The prototype system configuration

invalid, the listener comment display #22 "not-confident" is displayed. If the speaker's request is out of the system's domain knowledge, then the system displays a facial shrug and replies with "I cannot answer such a question." The relationships between conversational situations and facial displays are listed in Table 2.

## EXPERIMENT WITH THE PROTOTYPE SYSTEM

### Method

To examine the effect of facial displays in computer-human conversation, experiments were performed using the prototype system.

*Subjects.* The prototype system was tested on 32 volunteer subjects. Half were from the engineering staff of the computer product development division of Sony, while the remainder were university-level computer science students. The average age of the subjects was 26. All had experience of using computers, with an average career length of 7 years.

*Experiments.* Two experiments were prepared. In one experiment, called F, the subjects held a conversation with a system having facial displays. In the other experiment, called N, the subjects held a conversation with a system that displays short phrases instead of facial displays. The short phrases are four- or five-word sentences describing the corresponding facial displays. For example, instead of displaying #22 display "Not confident", "I am not confident" appears on the screen. The subjects were divided into two groups, FN and NF. As the names indicate, the subjects in the FN group first took experiment F then N, while those in the NF group first took N and then F.

In both experiments, the subjects were given the same conversation goal of enquiring about the functions and prices of Sony computer products. In each experiment, the subjects were requested to complete the conversation within 10 minutes.

*Measurements.* During the experiments, the number of occurrences of each facial display was counted. The conversation content was also evaluated based on the number of topics a subject visited. The results were scored using the following equation:

$$s = (3 + 2 * n + m) / t$$

where $s$, $n$, $m$, and $t$ are scores for conversation, number of topic shifts, number of successful answers, and the duration of the conversation. The subject's face and the system's face were videotaped during the experiments for later review. After completing both experiments, the subjects were asked to answer an inquiry sheet that asked the subjects to rate the qualities of speech recognition, facial displays, etc.

### Results

The results are shown in Takeuchi, Nagao, Color Plate 2, which plots the relative frequencies of facial displays (displays with no occurrence are omitted) and conversation scores (ACHIEVEMENT). Note that the conversation scores has a different scale from the others. Each experiment can be classified into one of two types, according to the characteristics of its result. The first type is "successful conversation" in which the conversation score is relatively high, and the displays "moderately confident", "beginning of story" appear more often. The second type is "not successful conversation" in the which conversation score is lower, and the displays "neutral" and "not confident" appear more often.

In Figure 2, the first experiments of the two groups are compared. It is clear that conversation with facial displays is more successful than conversation with short phrases. We can conclude that upon first contact facial displays clearly help conversation.

Table 2. Relation between conversational situations and facial displays

| Conversational situations | Facial displays |
|---|---|
| recognition failure | listener comment display #22 "not-confident" |
| syntactically invalid utterance | listener comment display #22 "not-confident" |
| many recognition candidates with close scores | listener comment display #21 "moderately-confident" |
| beginning of dialogue | listener comment display #18 "indication-of-attendance" |
| introduction to a topic | syntactic display #7 "beginning-of-story" |
| shift to another topic | syntactic displays #7 "end-of-story" and #9 "beginning-of-story" |
| answer "yes" | speaker display #14 "yes" |
| answer "no" | speaker display #15 "no" |
| out of the domain | speaker display #11 "facial shrug" |
| answer "yes" with emphasis | listener comment display #23 "yes" and syntactic display #3 "emphasizer" |
| violation of pragmatic constraints | listener comment display # 26 "incredulity" |
| reply to "thanks" | listener comment display #23 "yes" |
| ... | ... |

Figure 3 compares the overall results of both groups. The graph shows that the FN group is more successful than the NF group. Because the only difference between the two groups is the order in which experiments were conducted, we can conclude that early interaction with the system with facial displays improves later interaction.

Figure 4 compares the experiments with facial displays (1st of FN and 2nd of NF) and the experiments with short phrases (2nd of FN and 1st of NF). Contrary to our expectations, the results show relatively little influence by facial displays. This implies that the learning effect occurring over the first and second experiments is equal to the effect gained with facial displays. However, we believe that the effect gained with facial displays will be able to better the learning effect if the qualities of speech recognition and facial animation are improved.

## DISCUSSION AND FUTURE DIRECTION

The experiments show that facial displays are helpful especially upon first contact with the system. It is also shown that early interaction with facial displays improves successive interaction, even when there is no facial display. These results prove quantitatively that interfaces with facial displays reduce the mental barrier between the users and the computing systems.

Several premature facets of the prototype system fail to realize the potential advantages of a system with communicative facial displays. The system currently lacks lip synchronization and has a limited vocabulary. If these aspects are improved, the results would be much better. All subjects were relatively familiar with using computers. Experiments with non computer-literate users should also be done.

As a new research direction, the integration of more communication channels and modalities offers great promise. Among them, prosody information processing in speech recognition and speech synthesis are of special interest, as well as the recognition of a user's gestures and facial displays.

So far, conversation with computer systems has been over-regulated. This is because communication is done through limited channels. It is necessary to avoid information collision in these narrow channels. Multiple channels reduce the necessity of conversation regulation, so that new styles of conversation will appear, which have smaller granularity and high interruptibility, and which can invoke more spontaneous utterances. Such conversation is closer to our daily conversation with family and friends, and this

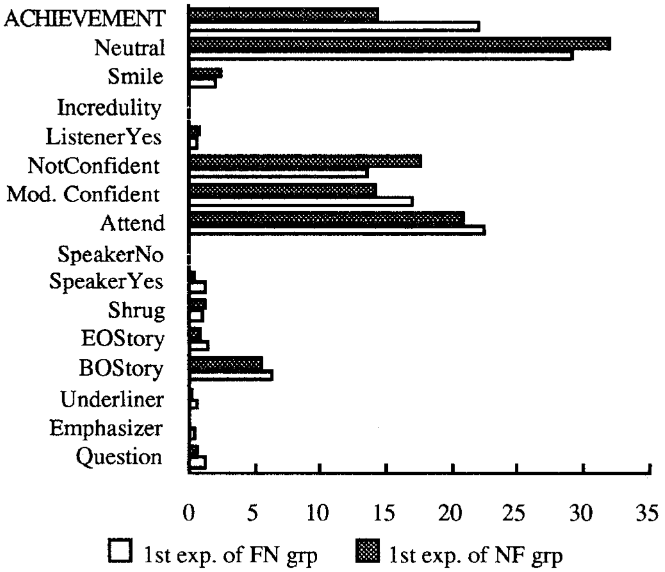


Figure 2. Comparison of the first experiments

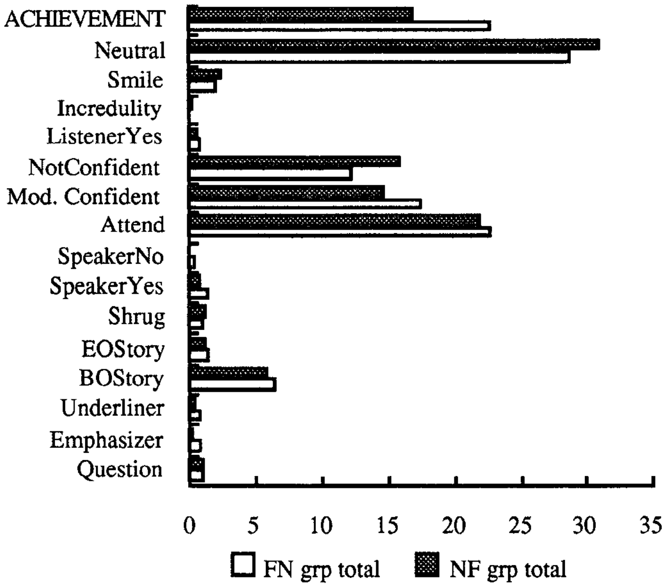


Figure 3. The effect of the first interaction

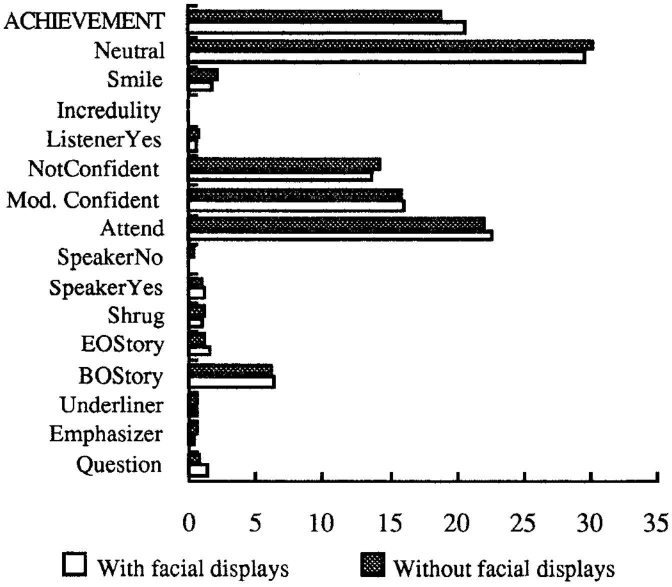


Figure 4. With or without facial displays

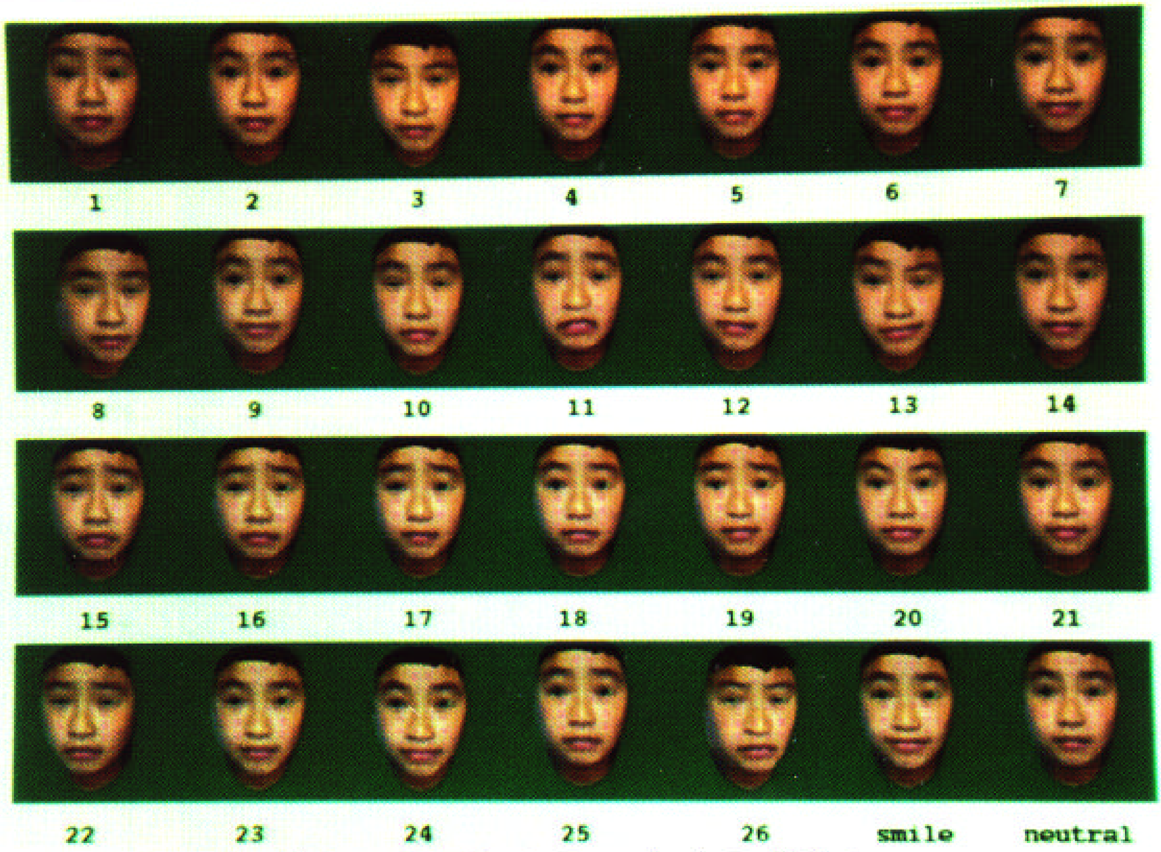
will further increase the user-friendliness of computers.

## ACKNOWLEDGMENTS

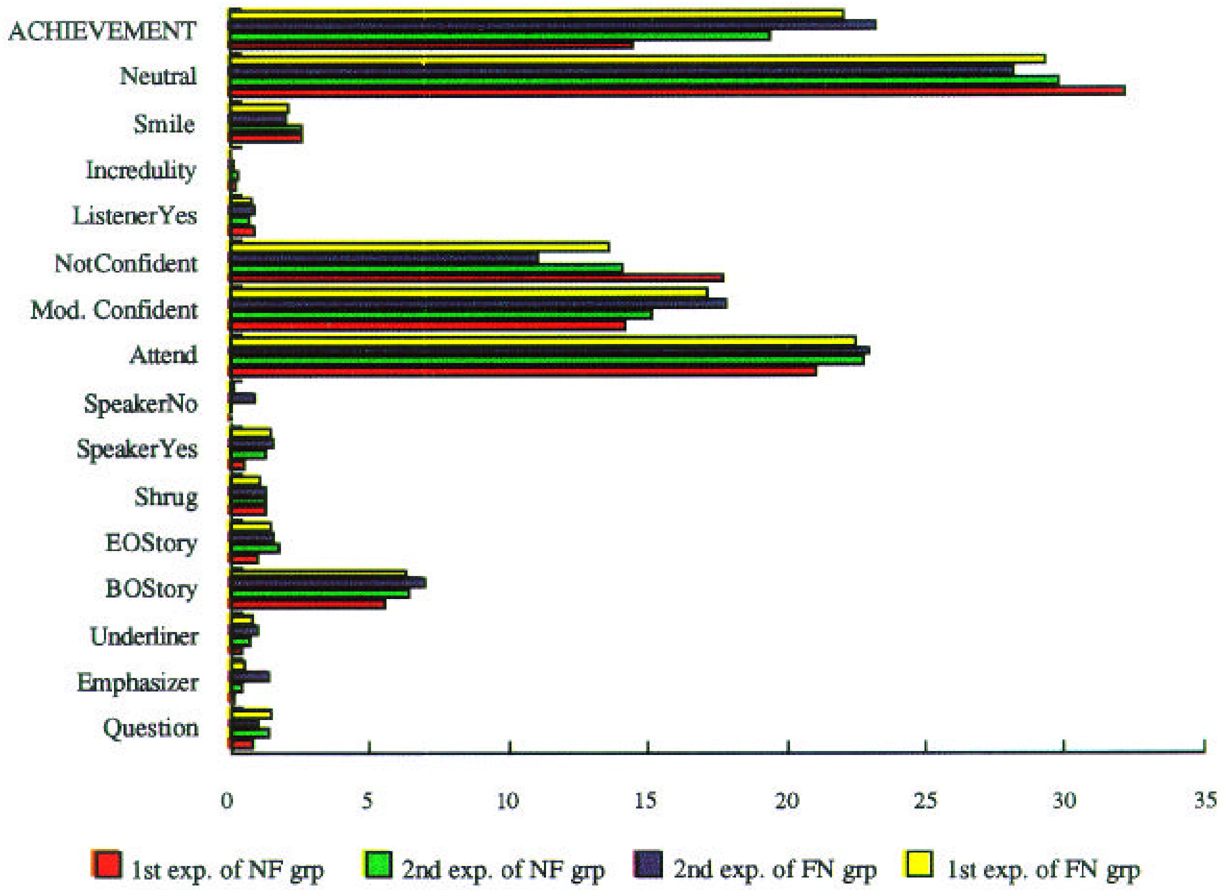
We would like to thank Alan Bond and Leslie Brothers for their suggestions and guidance in the early stage of the research. We also thank Steve Franks and Katunobu Itou for their contributions to implementing the prototype system. Special thanks to Keith Waters for permission to access his original animation system. Finally we thank Mario Tokoro and our colleagues at Sony CSL for their encouragement and discussion.

## REFERENCES

1. Blattner, M. *Multimedia and Multimodal User Interface Design: CHI'92 Tutorial Course Note 4*. ACM Press, 1992.
2. Chovil, N. *Communicative Functions of Facial Displays in Conversation.* Ph.D. Thesis, University of Victoria, 1989.
3. Darwin, C. The Expression of Emotion in Man and Animals. University of Chicago Press, Chicago, 1965.
4. Don, A. and Brennan, S. and Laurel, B. and Shneiderman, B. Anthropomorphism: from Eliza to Terminator 2, In Proc. CHI'92 Human Factors in Computing Systems (Monterey, May 3-7, 1992), ACM Press, pp. 67-70.
5. Don, A. and Oren, T. and Laurel, B. GUIDES 3.0, in Proc. CHI'91 Human Factors in Computing Systems (New Orleans, April 27-May 2, 1991), ACM Press, pp. 447-448.
6. Ekman, P. and Friesen, W. V. The repertoire of nonverbal behavior - categories, origins, usage, and coding, Semiotica 1 (1969), pp. 49-98.
7. Ekman, P. and Friesen, W. V. *Facial Action Coding System.* Consulting Psychologists Press, Palo Alto, California, 1978.
8. Ekman, P. and Friesen, W. V. *Unmasking the Face.* Consulting Psychologists Press, Inc., Palo Alto, California, 1984.
9. Fridlund, A. J. and Gilbert, A. N. Emotions and facial expression, Science, 230 (1985), pp. 607-608.
10. Hindus, D. and Brennan, S. *Conversational Paradigms in User Interfaces: CHI'92 Tutorial Course Note 11.* ACM Press, 1992.
11. Itou, K. and Hayamizu, S. and Tanaka, H. Continuous speech recognition by context-dependent phonetic HMM and an efficient algorithm for finding N-best sentence hypotheses, in Proc. ICASSP'92, IEEE Press, pp. I 21-I 24.
12. Nagao, K. A preferential constraint satisfaction technique for natural language analysis, in Proc. ECAI-92, (1992), pp. 523-527.
13. Nagao, K. and Osawa, E. *A Logic-Based Approach to Plan Recognition and Belief Revision.* Tech. Report. SCSL-TR-92-007, Sony Computer Science Laboratory, Inc., Tokyo, 1992.
14. Perret, D. I. et al. Neurones responsive to faces in the temporal cortex: studies of functional organization sensitivity and relation to perception. Human Neurobiology, 3 (1984) 197-208.
15. Sherer, K. R. The functions of nonverbal signs in conversation, in *The Social and Psychological Contexts of Language*, St. Clair, R. N. and Giles, H. (Eds.), Lawrence Erlbaum, Hillsdale, NJ, 1980, pp. 225-244.
16. Takeuchi, A. and Franks, S. *A Rapid Face Construction Lab.* Tech. Report. SCSL-TR-92-010, Sony Computer Science Laboratory, Inc., Tokyo, 1992.
17. Tomita, M. An efficient augmented-context-free parsing algorithm, Computational Linguistics, 13 (1987), pp. 31-46.
18. Waters, K. A muscle model for animating three-dimensional facial expression, in Computer Graphics 21, 4 (July 1987), 17-24.



**Takeuchi**, Nagao, Color Plate 1. Communicative Facial Displays


ACHIEVEMENT
Neutral
Smile
Incredulity
ListenerYes
NotConfident
Mod. Confident
Attend
SpeakerNo
SpeakerYes
Shrug
EOStory
BOStory
Underliner
Emphasizer
Question
🟥 1st exp. of NF grp
🟩 2nd exp. of NF grp
🟦 2nd exp. of FN grp
🟨 1st exp. of FN grp